DOCUMENT RESUME

ED 076 706 TM 002 715

AUTHOR Touq, M. S.; And Others
TITLE Criterion-Referenced Validity of Student Ratings of Instructors.
PUB DATE 73
NOTE 9p.; Paper presented at annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS College Students; Observation; Rating Scales; *Student Opinion; Teacher Behavior; *Teacher Rating; Technical Reports; *Validity; Verbal Communication

ABSTRACT

The purpose of this research was to assess the criterion-referenced validity of student ratings of instructors. A total of 480 undergraduates rated their instructors using a special rating scale designed to parallel the Flanders Interaction Analysis Categories. Expert observers also rated the instructors using the standard form of the Flanders Categories. Mean student ratings for instructors were correlated with expert observers' scores. Significant correlations were found between ratings for four categories. These results were interpreted as revealing some criterion-referenced validity for student ratings. (Author)

Criterion-Referenced Validity of Student Ratings of Instructors

M. S. Touq

University of Jordan

J. F. Feldhusen and J. Halstead

Purdue University

Research on the reliability of student ratings of instruction indicates that students are indeed reliable raters of their instructors. Reliability coefficients range from moderately positive to high positive correlations (McKeachie, 1969). However, very little research has been reported on the validity of student ratings of instruction.

Most researchers and users of student ratings of instruction are satisfied with face validity of the instruments if the content of items seems to focus on significant aspects of instruction (Remmers, 1963). Studies of the construct validity of student rating forms through factor analysis have been only moderately successful in identifying replicable and interpretible components of teacher behavior (Derry, 1972). A number of researchers have also assessed the concurrent or predictive validity of student ratings of instruction by correlating student ratings with ratings of the same instructors

---

A paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

by alumni (Drucker and Remmers, 1951), colleagues (Guthrie, 1954; Maslow and Zimmerman, 1956), and supervisors (Costin et. al. 1971; Hayes, 1971). Substantial agreement among different groups has been found.

Perhaps the ideal way to deal with the validity problem and to assess the accuracy of students' ratings was proposed by Halstead, Feldhusen and McDaniel (1970). They suggested that rating of instruction be done by expert observers and the results compared with student ratings. Like the studies of concurrent and predictive validity, this is an evaluation of criterion-referenced validity. This approach was used in the present study. The questions were stated as follows: Are student and teacher verbal behaviors as observed by professional observers correlated with ratings of these same behaviors by the students themselves? Are there significant differences between student and expert observers in the amount of each type of behavior observed?

Methods

Subjects

Eighteen instructors, twelve males and six females, and 488 undergraduate students enrolled in eight educational psychology classes, eight general psychology classes, and two sociology classes were the subjects for this study. These sections were taught by five instructors and fourteen graduate

and teaching assistants. Approximately one third of the students were males and two thirds were females. Students ranged from freshman to seniors in college. The number of students in classes ranged from 18 to 44 with a mean of 27.3.

Procedures:

Flanders Interaction Analysis Categories (FIAC; Flanders, 1970) was used to assess student-teacher verbal interactions. Two trained observers visited the classes and observed and recorded the interactions. Inter-rater reliability was .85.

The following teacher behaviors and interactions were assessed: (1) acceptance of feelings, (2) praise and encouragement, (3) use of student ideas, (4) asking questions, (5) lecturing, (6) giving directions, (7) criticizing, (8) student talk - response, and (9) student talk - initiation.

To obtain student ratings of teacher behavior and student-teacher interactions, an Interaction Analysis Questionnaire (IAQ) was developed and administered to the students (Touq, 1972). This questionnaire consists of nine items representing student and teacher verbal behaviors parallel to the first nine categories of the FIAC (Flanders, 1970). Test-retest reliability was found to be .75.

Scores on both the FIAC and the IAQ were percentages of classroom time spent in each of the nine types of behavior.

Frequencies of the FIAC were then correlated with student ratings of instructors on the IAQ for the parallel categories. Alpha was set at .10. Differences between means for each category on FIAC and IAQ were evaluated with a t test for correlated means with alpha equals .05.

Results

Table 1 shows the means of student ratings of classroom interaction activities for all the classes involved in this study and the assessments of the same activities utilizing the FIAC. Table 1 also gives the correlations between the IAQ mean scores and the FIAC scores. Four correlations out of nine were significant (.43. .49, .44, and .61) with a fifth correlation approaching significance (.36). "Accepting feelings" on the FIAC had a significant and negative correlation with the same category on the IAQ (-.43). "Praising and encouraging" on the FIAC had a significant and positive correlation with the same category on the IAQ (.49). "Lecturing" on the FIAC had a significant and positive correlation with the same category on the IAQ (.44). "Student talk - initiation" on the FIAC had a significant and positive correlation with the same category on the IAQ (.61). Correlation between "Student talk - response" on the FIAC and the IAQ also approached significance (.36).

Differences between the means for each parallel category on FIAC and IAQ were tested using the t test for correlated

means (Winer, 1971) and an alpha level of .05. The results indicate that the differences were significant for seven out of the nine means. These were "accepting feelings" (t = 14.83), "praising or encouraging" (t = 14.97), "accepting ideas" (t = 8.69), "lecturing" (t = 7.06), "giving directions" (t = 2.29), "criticizing or justifying authority" (t = 3.59), "student talk - response" (t = 9.05). The differences between means of the FIAC and the IAQ were not significant for "asking questions" (t = 1.54) and "student talk - initiation" (t = 1.45).

Discussion

The first question asked in this research was: Are student and teacher verbal behaviors as observed by professional observers correlated with ratings of these same teacher behaviors by the students themselves? The answer is affirmative. Three significant and positive correlations were found. One, the correlation between FIAC and IAQ "student talk - initiation," was .61. The other significant ones were "praising or encouraging" (r = .49) and "lecturing" (r = .44). The correlation for category 1, "accepting feelings" (-.43) was significant and negative.

The second question was: Are there significant differences in the amount of each type of behavior observed between student and expert observers? Significant differences were found for "accepting feelings", "praising or encouraging","accepting ideas", "lecturing", "giving directions", "criticizing or justifying

authority", student talk - response. The differences were not significant for "asking questions" and student talk - initiation. The means of the IAQ categories were all larger than the means of the FIAC categories except for category five (lecturing) where the mean of the FIAC was larger than the mean of the IAQ.

The correlations found in this study indicate some agreements between students and expert observers with regard to instructors' classroom behaviors. Thus, there is moderate support for the criterion-referenced validity of student ratings of instruction. Of particular significance is students' perceptions of their own behavior. Students were most accurate in assessing their own initiated talk in classroom. The correlations with expert observers was .60 and there was no difference between the FIAC and IAQ means. The fact that the correlation for "student talk - response" was not significant and the difference between FIAC and IAQ means was so great might be due to some confusion on the part of the students in making differentiation between initiated talk and talk in response to a question.

Of particular interest is the significant negative correlation for Category 1, "accepting feelings", between the FIAC and the IAQ. This is coupled with the large difference between means. Students see much more of this behavior than observers. Perhaps the students are rating on the basis of out-of-class teacher behaviors. But this still leaves open the question of the negative correlation.

It is possible to speculate that the teacher who shows little acceptance of student feelings in class shows much in personal conferences in his office. Conversely the teacher who demonstrates acceptance of student feelings in class shows no such acceptance in personal contacts and thus is rated down by students.

A number of researchers have indicated that student ratings are valid when they are evaluated against different criteria such as alumni, colleagues, and supervisors (Drucker and Remmers, 1951; Guthrie, 1954; Costin, et. al. 1971; Maslow and Zimmerman, 1956; Clark and Blackburn, 1971; and Hayes, 1971). Thus, the results of this study add more support for the findings of these researchers. However, the approach of this study to criterion-referenced validity is unique and probably more important than the other approaches because outside professional observers have no personal stake in the educational process that might bias their ratings and because they are knowledgeable about instruction.

Higher correlations might be obtained if there was some assurance that the students understood the specific behaviors they were rating. The subjects of this study were not previously exposed to either the FIAC or its parallel form the IAQ. Training students on these scales might increase the accuracy of their IAQ ratings. Halstead, Feldhusen and McDaniel (1970) proposed such a procedure. Halstead (1972) carried out research which was partially successful in improving the reliability of student rating through training the students in the rating procedures.

Summary

The purpose of this research was to assess the criterion-referenced validity of student ratings of instructors. A total of 480 undergraduates rated their instructors using a special rating scale designed to parallel the Flanders Interaction Analysis Categories. Expert observers also rated the instructors using the standard form of the Flanders Categories. Mean student ratings for instructors were correlated with expert observers' scores. Significant correlations were found between ratings for four categories. These results were interpreted as revealing some criterion-referenced validity for student ratings.

Table 1

Means and Standard Deviations
For FIAC and IAQ Categories

Category	FIAC Mean	FIAC Standard Deviation	IAQ Mean	IAQ Standard Deviation	Correlation
(1) Accepting feelings	0.08	.17	11.70	3.70	-.43*
(2) Praising or encouraging	1.69	1.18	8.27	2.01	.49*
(3) Accepting ideas	3.37	2.39	8.28	2.75	.16
(4) Asking questions	9.98	13.97	8.96	3.13	.01
(5) Lecturing	59.52	25.62	36.27	12.29	.44*
(6) Giving directions	1.23	1.34	3.98	5.99	.12
(7) Criticizing	0.42	1.48	1.66	1.41	-.08
(8) Student talk - response	4.37	3.55	11.44	3.11	.36
(9) Student talk - initiated	14.54	17.32	9.37	3.10	.61*

*Significant

References

Clark, M. J. and Blackburn, R. T. "Assessment of Faculty Performance: (1) Methodology, and (2) Some Correlates Between Self, Colleague, Student, and Administrative Ratings," University of Michigan, Unpublished Technical Report, 1971.

Costin, F., Greenough, W. T. and Menges, R. J. "Student Ratings of College Teaching: Reliability, Validity and Usefulness." Review of Educational Research, 1971, 41, 511-535.

Derry, J. D. "Factor Analysis of Instructor Ratings." Unpublished Document, Measurement and Research Center, Purdue University, 1972.

Druckers, A. J. and Remmers, H. H. "Do Alumni and students differ in their attitudes toward instructors?" Journal of Educational Psychology, 1951, 42, 129-143.

Flanders, N. A. Analyzing Teaching Behavior. Reading, Mass.: Addison-Wesley, 1970.

Guthrie, E. R. The Evaluation of Teaching: A progress report. Seattle: University of Washington, 1954.

Halstead, J. S. "Student Ratings of College Classroom Verbal Interaction As Related To Ratings of Instructor Teaching Effectiveness." Unpublished Doctoral Dissertation, Purdue University, 1972.

Halstead, J. S., Feldhusen, J. F., and McDaniel, D. D. "Models for Research on Ratings of Courses and Instructors." Proceedings of 78th Annual Convention, APA, 1970, 625-626.

Hayes, J. R. "Research, Teaching, and Faculty Fate." Science, 1971, 172, 227-230.

Maslow, A. H. and Zimmerman, W. "College Teaching Ability, Scholarly Activity and Personality." Journal of Educational Psychology, 1956, 47, 185-189.

McKeachie, W. J. Teaching Tips: A Guidebook for the Beginning College Teacher. Lexington, Mass.: D.C. Heath, 1969.

Remmers, H. H. "Rating Methods in Research on Teaching." In N. L. Gage (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally and Company, 1963, 329-378.

Touq, M. S. "The Relationship Between Student Participation in Classroom Discussion and Student Ratings of Instructors at the College Level." Unpublisned Doctoral Dissertation, Purdue University, 1972.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, 1971.